Pose Estimation on Antelopes in the Wild

Shaan Chanchani
Department of Engineering
Purdue University
West Lafayette, USA
schancha@purdue.edu

Aryan Khanolkar
Department of Computer Science
Purdue University
West Lafayette, USA
akhanol@purdue.edu

Claire Kim
Department of Engineering
Purdue University
West Lafayette, USA
kim3386@purdue.edu

Josh Mansky
Department of Engineering
Purdue University
West Lafayette, USA
jmansky@purdue.edu

Medhashree Parhy
Department of Computer Science
Purdue University
West Lafayette, USA
mparhy@purdue.edu

Armaan Sayyad
Department of Data Science
Purdue University
West Lafayette, USA
asayyad@purdue.edu

Parth Thakre
Department of Computer Science
Purdue University
West Lafayette, USA
pthakre@purdue.edu

[1] ***Abstract*—Our research explores two major factors that impact the generalizability of keypoint estimation models to more difficult data. Specifically, we focus on in-the-wild antelope images captured from motion-triggered camera traps in Senegal, West Africa.**

The first major factor we will focus on is how our model's performance changes based on varied subsets of training data. To do so, we leverage the AP-10k [1] dataset to explore different training strategies, such as whether training on a small subset of visually similar species helps the model generalize better.

Our second area of exploration is testing our keypoint labeling scheme, which is more rigorous than the labeling scheme of the AP-10K dataset. Using our data, we created keypoint labels based on several different definitions – some are visually distinct, while others are more biologically correct. By training and testing with these different keypoint definitions, we want to explore what kind of keypoint definition helps the model generalize better. Our results improve keypoint detection for animals and, in a broader sense, contribute to the abundance estimation of animals in the wild.

I. Introduction

This study aims to improve pose estimation accuracy on antelope images captured from motion-triggered camera traps in Senegal, West Africa, as part of research on chimp prey abundance. We trained RTMPose [2] models, selected for their efficiency over HRNet [3], using four subsets of the AP-10k [1] dataset maintaining equal dataset size and architecture for each model. Tested on 100 consistent antelope images, the results of the model training highlighted how taxonomically and visually similar species improve model generalization. Additionally, we developed and evaluated a custom keypoint labeling scheme, refining definitions to enhance labeling consistency, which supports better model generalization across low-quality, real-world images. This approach demonstrates how dataset composition and precise keypoint definition scan boost pose estimation accuracy, aiding behavioral studies when labeled data for specific species is scarce.

II. Background

Pose estimation involves identifying keypoints (joints or landmarks) on a subject to understand their configuration and infer posture or movement. Outputs of pose estimation models are often represented as skeleton-like stick figures superimposed on images and video frames. We choose antelopes as the starting point because they usually have more rigid body than primates; i.e. their movements are more constrained by their physical skeleton. Pose estimation also offers valuable insights into prey-predator interactions by analyzing antelope postures and behaviors during predator encounters to understand survival strategies. Furthermore, pose estimation can contribute to livestock management by enabling the diagnosis of disease-related postures and facilitating comprehensive herd health monitoring. Some challenges that we come across when applying pose estimation on our antelope images from Senegal include limited annotated antelope datasets, occlusions in group settings, diverse environmental conditions, and fine-grained species anatomical differences. The dataset we use, AP-10K [1], is a comprehensive resource with 10,000 labeled images and 50,000 unlabeled images across 53 species. The dataset contains 200 specific antelope images.

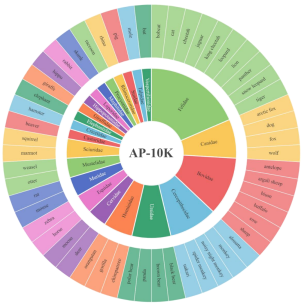


Fig. 1. AP-10K dataset composition

Senegal Antelope dataset comprises camera trap images currently in the process of being labeled. We ran MegaDetector

[1] Authors ordered in alphabetical order by last name. With thanks to Haoyu Chen and Dr. Amy Reibman

[4] on all unlabeled images first, which generates bounding boxes of potential animals; however, the detector is far from perfect, and the results often contains many false detections. We applied a few-shot-learning classifier [5] to select images that most likely to be antelopes. We then manually pick out actual antelopes, to construct our own antelope dataset. The antelope species we have observed includes cape bushbuck (*Tragelaphus sylvaticus*), red-flanked duiker (*Cephalophus rufilatus*), oribi (*Ourebia ourebi*), roan antelope (*Hippotragus equinus*), hartebeest (*Alcelaphus buselaphus*), giant eland (*Taurotragus derbianus*). In search of a model to do pose estimation on our low quality data, we identified two reputable animal pose models. These were HRNet [3] and RTMPose [2], both known for their strong performance when trained on the animal pose datasets such as AP-10K [1]. When these models were tested on our images, RTMPose [2] was selected over HRNet [3] due to its superior performance particularly in challenging imaging conditions that involved occlusions. An example of this is th first image comparison in Figure 2, in which the occlusio caused by a tree disrupts HRNet's [3] pose estimation, as i assumes the tree to be the hind-legs of the antelope. RTMPos [2], on the other hand, demonstrated a notable ability to mor accurately identify keypoints across various scenarios (such a occlusions, or difficult visual contexts), making it especiall suitable for the lower quality data at hand.

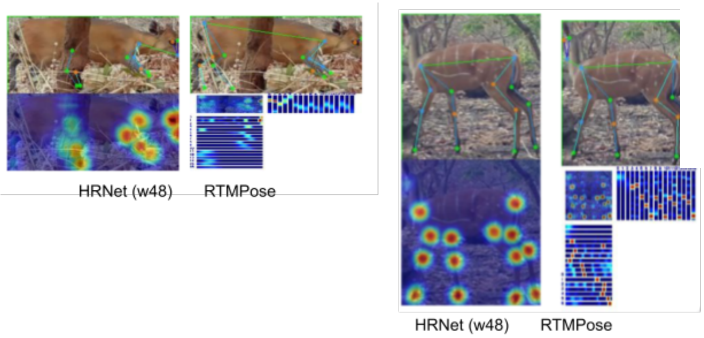


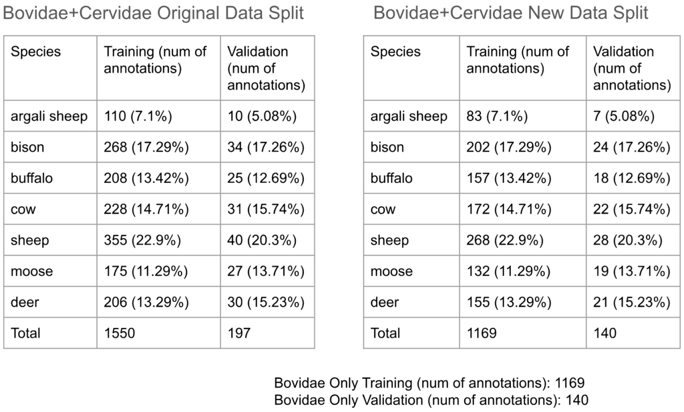
Fig. 2. Model Labeling Results

III. Progress Made

We focused on three main parts of the project. Firstly, we decided to do subset-based training, to explore the effects of the training data on model performance. Then, to further improve performance, we decided to incorporate low-quality into the training data by adding labeled Senegal antelope data. To add labels to the data, we needed to create a definition of keypoints to ensure both consistency across our annotators and robustness for every keypoint. The third portion of the project was creating a method to identify similar-looking species to aid in further research into the subset-training by adding in more similar-looking species.

A. Subset-Based Training

The first aspect of our research was improving the pose estimation model by expanding the training dataset beyond just antelopes. Our hypothesis was that adding similar species to antelopes during the training process could help the model generalize better and improve accuracy. We used a taxonomically-based similarity measure, by choosing to add in species based on the taxonomic family that they belong to. To test this, we decided to train the RTMPose [2] model on three different subsets: Bovidae, Bovidae + Cervidae, and the entire AP-10K [1] dataset. Bovidae is the family that antelopes belong to, which includes argali sheep, bison, buffalo, cows, and sheep. Cervidae includes moose and deer, which we added because they resemble antelopes. To ensure that the size of the dataset was not a factor in our experiment, all datasets were reduced to match the size of the Bovidae set (the smallest of the three sets). This reduction was done by percentage, maintaining the original distribution of animals while decreasing the overall size, as seen in Figure 3.



Bovidae+Cervidae Original Data Split

Species	Training (num of annotations)	Validation (num of annotations)
argali sheep	110 (7.1%)	10 (5.08%)
bison	268 (17.29%)	34 (17.26%)
buffalo	208 (13.42%)	25 (12.69%)
cow	228 (14.71%)	31 (15.74%)
sheep	355 (22.9%)	40 (20.3%)
moose	175 (11.29%)	27 (13.71%)
deer	206 (13.29%)	30 (15.23%)
Total	1550	197

Bovidae+Cervidae New Data Split

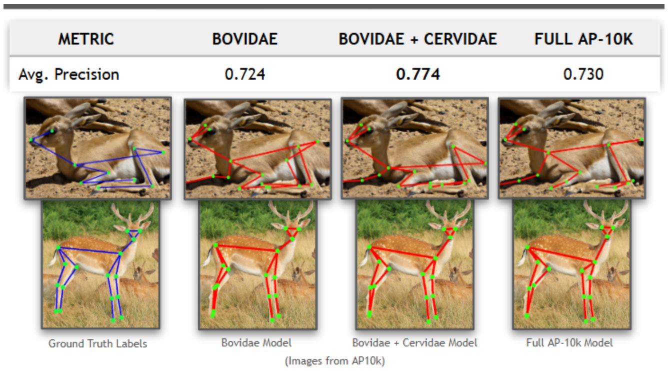
Species	Training (num of annotations)	Validation (num of annotations)
argali sheep	83 (7.1%)	7 (5.08%)
bison	202 (17.29%)	24 (17.26%)
buffalo	157 (13.42%)	18 (12.69%)
cow	172 (14.71%)	22 (15.74%)
sheep	268 (22.9%)	28 (20.3%)
moose	132 (11.29%)	19 (13.71%)
deer	155 (13.29%)	21 (15.23%)
Total	1169	140

Bovidae Only Training (num of annotations): 1169
Bovidae Only Validation (num of annotations): 140

Fig. 3. Model Labeling Results

All datasets were created using customized parsing scripts available on our project GitHub [6] (https://github.com/VIP-VAA-Fall24/MD_doc). Due to a lack of images (200), we excluded antelopes from the training and validation sets. Testing was conducted on 100 set-aside antelope images, consistent across all model tests. Additionally, there are 100 set-aside antelope images available for fine-tuning the models, which can be done in the future to further enhance performance.

To summarize our results across the models with varied training sets, we create Figure 4, which provides both quantitative and qualitative result. During the testing process, all of our metrics followed the same trends, so we decided to just focus on average precison. In Figure 4, for a visual of the quantitative results, we used the three models to inference on the same images, to get a comparision to the ground truth labels from the AP10k [1] dataset. We found that adding Cervidae images improved average precision compared to the only Bovidae model by 5% as seen in Figure 3. One the other hand, incorporating all species in the AP-10K [1] dataset only improved average precision by 0.6%. This suggests a correlation between adding similar-looking images to the



METRIC	BOVIDAE	BOVIDAE + CERVIDAE	FULL AP-10K
Avg. Precision	0.724	0.774	0.730

Fig. 4. Precision across models

training set and improved performance. This correlation could be further explored by varying the training set with similar images and adjusting the species distribution (e.g., changing the proportion of sheep from 22.9% as seen in Figure 3 to around 10%) to study its effects.

We can also further explore the subset-training by using the species-similarity method we are working on to add in more similarly looking species to the training set.

B. Keypoint Definitions

The primary objective was to improve the pose estimation model's accuracy on more challenging data. The data in AP-10K is high-quality and unoccluded, so we decided to include Senegal camera-trap antelope images to increase model performance. The AP-10K dataset lacked a publicly available labeling scheme. As such, in the process of labeling the Senegal dataset, we would not be able to replicate the AP10k data accurately.

To address this, we decided to develop a custom labeling scheme to ensure both consistency across our labelers and robustness for every keypoint. This would give us a well-defined labeling scheme, that we could then check annotations against. We took the opportunity to create a definition, to iteratively adjust more difficult keypoints. We encountered issues with maintaining consistency for the hip, neck, and shoulder keypoints, since there is no significant skeletal point for either point. Since the points are located on a general area of the skin, the labeling was done based on intuition, which made it difficult to maintain consistency.

Our first iteration of designs, visualized in Figure 5, was to use a crossection between the base of the neck, and a line parallel to the top of the neck (visualized in red) coming from the base of the skull. This design was very complicated, and difficult for labelers to understand. Upon practical trials, we found that for some poses of the antelope, it was very difficult to find the lines described. Our next design worked to simplify this by simply marking the midpoint of the base of the neck (marked by the dotted line). This did work practically, but we still had inconsistencies of the point due the midpoint being marked by eye. To resolve this, we moved to our final design, in which we chose to label two distinct points and then interpolate the originally defined keypoint. Specifically, instead of labeling the neck keypoint in the middle of the neck, we labeled the throat and wither keypoints, with the neck keypoint being calculated as the midpoint during preprocessing, which increased robustness of the point.

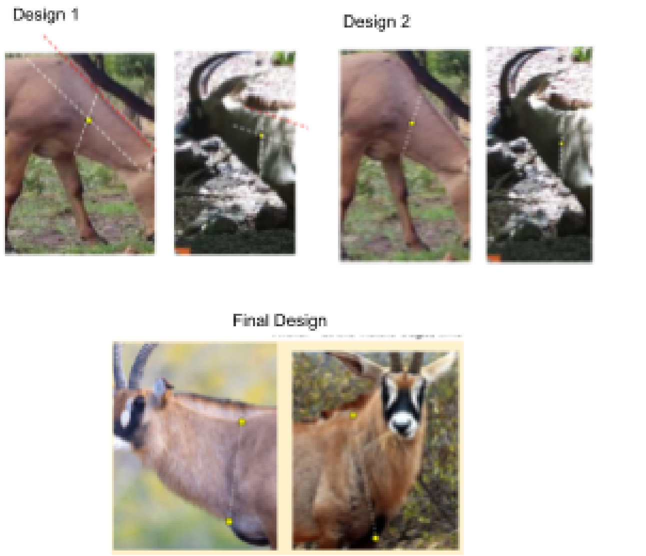


Fig. 5. Neck keypoint iterations

A similar process was applied to the shoulder, by labeling two points on the front legs, and then finding the midpoint via preprocessing. The hip keypoint was interpolated from the root-of-tail keypoint and a keypoint marking the edge between the leg and the torso. THis largely resolved our issue of using intuition to place a point on the general hip area. We also faced issues with the root-of-tail keypoint, so after multiple iterations, we decided to label the base of the tail and added images from various angles to further clarify. To ensure the final keypoint definitions were effective, we used the definitions on a few images, and qualitatively compared the labels against each other to ensure consistency. Our keypoint definitions can be found on our project Github [6] (https://github.com/VIP-VAA-Fall24/MD_doc/blob/main/Definitions+Documentation/Keypoint_Definitions-10_28_2024.pdf)

C. COCO Format AP-10k

We created a quick guide for the COCO data format, which is the data format of the AP-10K annotations. This guide identifies important keys and flags such as the visibility flag for the key points, indicating whether a key point is visible, partially occluded, or fully occluded, and so not labeled. Creating this guide was crucial as it allows us to understand the data format we needed to attain after labeling our keypoints for model training and testing. Along with COCO being the primary format used for our input data for the models, understanding all the details such as the keypoint flags is crucial as it allows us to experiment with different labeling processes. Our COCO data guide can be found on our project Github [6] (https://github.com/VIP-VAA-Fall24/MD_ doc/blob/main/Definitions+Documentation/AP10K_COCO_JSON_Format.pdf)

D. Labeling Effort

After evaluating several major data annotation platforms, we selected Label Studio as it best suited our requirements for collaborative keypoint labeling. During the evaluation, we looked to find the most flexible and easy-to-use platforms, allowing our team to make changes to keypoint labeling stratergy and process easily. Currently, we have installed and configured Label Studio with a database on our shared GPU server, along with a startup script enabling concurrent labeling sessions. The software exports labeled keypoints in JSON files and allows us to distribute our dataset of Senegalese antelope images such that each image receives at least two independent labels to resolve potential discrepancies. To accommodate our custom labeling scheme where certain keypoints are derived from the midpoint of two reference points, we developed post-processing scripts that compute these interpolated keypoints from the directly labeled reference points. This setup enables efficient collaborative labeling, making anyone's labels instantaneously accessible to anyone else on the team.

E. Species Similarity

In addition to our taxonomy-based approach, we explored using visual similarity between species to inform dataset selection for pose estimation training. While taxonomic relationships provide a natural way to group similar animals, we hypothesized that visual similarity might be more relevant for pose estimation performance. This led us to develop a methodology for quantitatively measuring species similarity to guide data selection. To accurately compare species proportions and anatomy, we needed standardized poses across our sample images. Side-view poses were identified as ideal since they provide a consistent perspective for measuring relative proportions and anatomical features. This standardization is crucial for creating reliable similarity metrics, as variations in pose and camera angle would introduce unwanted variance in our measurements.

Our initial approach used a keypoint-based heuristic to identify side-view poses by analyzing the relative positions of keypoints in images. The algorithm looked for images where the nose/eyes were the rightmost visible points and the root of tail/hips were the leftmost visible points. In its current state, our implementation isn't able to extract a sufficient amount of images to form a reliable sample size for each four-legged species in AP-10k. While this method could theoretically be expanded to capture the opposite orientation (left-facing poses) through image flipping, doing so would double the number of false positives in our results. The current approach already yields a considerable amount of false positives from animals that are photographed walking diagonally to the camera – these poses technically meet our criteria despite not being true side-views. Additionally, even among the true side-view images we capture, the current approach provides no mechanism to assess their relative quality. We can't automatically distinguish between a perfect side-view and one that's slightly angled, which makes it difficult to select the most ideal samples for our analysis.

To address these limitations, we have proposed a revised framework combining depth estimation and segmentation. This proposed method would leverage the existing bounding box annotations in the AP10-k dataset, using Apple's Depth Pro model for initial depth estimation. While this model may not provide perfectly accurate depth maps, its estimates should be sufficient for our specific use case of identifying true side-view poses. The pipeline would then use Meta's SAM2 segmentation model with bounding box prompts to generate precise masks of each animal instance. SAM2 has been demonstrated to generalize exceptionally well to animal segmentation across different species and environments, making it an ideal off-the-shelf solution for our needs. By combining the depth maps with segmentation masks, we propose two potential approaches for side-view detection: analyzing the variance in masked depth values or evaluating depth consistency around keypoint regions. While either approach has not yet been implemented or tested, it theoretically offers more robust detection of true side-views and the ability to rank images by their side-view quality.

Once we obtain high-quality side-view samples, we plan to create vector embeddings containing the ratios of distances between all possible keypoint combinations. These embeddings would be averaged across samples for each species, allowing us to use cosine similarity between species vectors as a quantitative similarity metric. This approach will allow us to systematically test how visual similarity correlates with pose estimation performance, providing insights into optimal dataset composition for training models on new species.

IV. Future Work

Some of our future work will be to build upon the progress we have made, but also experiment new methods outside of the ones we used so far. Firstly, additional work on using taxonomic similarity to improve pose-estimation can be done. So far, we merely took the given distribution of the AP10k dataset and scaled accordingly. Further adjusting the species distribution, (for example training with less sheep and more deer) can allow us to further understand the factors that influence the accuracy and generalization of the model.

Further work in morphological species similarity includes developing and implementing a robust metric integrating multiple factors such as; average species limb ratios, muscularity, range of motion.

This work will help us evaluate the key point estimation model's performance when trained on visually and taxonomically similar species. A crucial component of this research involves the creation of a specialized dataset of camera trap antelope images with corresponding keypoint annotations. By producing this dataset we aim to fine-tune and test the keypoint estimation model across various dataset subsets, with an emphasis on exploring how the distribution of species with respect to these subsets impact model performance.

References

[1] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao, “Ap-10k: A benchmark for animal pose estimation in the wild,” *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[2] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen, “Rtmpose: Real-time multi-person pose estimation based on mmpose,” *arXiv*, 2023.

[3] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang, “Deep high-resolution representation learning for human pose estimation,” *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5686–5696, 2019.

[4] Sara Beery, Dan Morris, and Siyu Yang, “Efficient pipeline for camera trap image review,” *arXiv preprint arXiv:1907.06772*, 2019.

[5] Haoyu Chen, Stacy Lindshield, Papa Ibnou Ndiaye, Yaya Hamady Ndiaye, Jill D. Pruetz, and Amy R. Reibman, “Applying few-shot learning for in-the-wild camera-trap species classification,” *AI*, vol. 4, no. 3, pp. 574–597, 2023.

[6] Shaan Chanchani, Aryan Khanolkar, Claire Kim, Josh Mansky, Medhashree Parhy, Armaan Sayyad, and Parth Thakre, “VIP-VAA-Fall24,” https://github.com/VIP-VAA-Fall24/MD_doc/tree/main, 2024.